The Data Access and Dissemination Systems Office: An Overview

Version 3

Date: April 22, 2004

This page has been left blank intentionally.

Table of Contents

Index of Tables

Index of Figures

1 Introduction

This paper provides an overview of the Data Access and Dissemination Office (the Office) at the United States Census Bureau. The intent of this paper is to describe the Office, its services and resources. The key components of this paper are a description of the organizational context of the Office within the Census Bureau, a detailed description of the services currently provided and the systems the Office has developed and now maintains.

This paper supplements other materials and information that describe the goals for a future integrated dissemination system that will integrate dissemination functions currently spread across a number of Census Bureau organizations.

1.1 Census Bureau

The Census Bureau is an agency of the Department of Commerce and is one of several principal agencies in the Federal Statistical System.[1] As such, the agency is charged with providing ongoing measures of the United States population and its economy. In doing so, it supports the economic and political foundations of the country and provides a critical service to the Nation. Therefore, reliable collection, preparation, and dissemination of demographic and economic statistics are vital to the overall mission and strategic goals of the agency.

Relevant, accurate and timely statistics support decision making at every level of government. They allow for the administration and equitable funding of many federal, state and local programs, and are used to fulfill constitutional and legislative mandates. One example is Article 2, Section 1 of the Constitution that requires an enumeration of the population every 10 years to apportion the seats in the House of Representatives among the states. States use additional data to redraw their legislative districts.

Statistics collected and disseminated by the Census Bureau are used by legislators, policy makers, educators, planners, businesses, non-profit organizations and the general public on a daily basis and for numerous purposes. Other important Census Bureau measures are part of the fabric of today’s information age, including monthly housing starts, new home sales, the monthly unemployment rate, the balance of trade, the poverty rate, construction spending and more.

1.2 Organizational Structure

The Census Bureau is organized into directorates, program areas based on function or subject matter as they relate to the design, collection, processing and dissemination of censuses, surveys, projections and estimates. In addition, there are other areas that provide key administrative and infrastructure support. Within directorates, smaller work

[1] See http://www.fedstats.gov/agencies/agencies.html for more information on the Federal Statistical System and access to other web sites.

units referred to as divisions are established based on the more detailed function or subject matter.[2] Currently, the Office provides services to four key directorates:

> **Communications Directorate:** Responsible for official news releases, press releases and other notifications of product and data releases; responding to and providing services for the media; and official liaison to the Congress.
>
> **Decennial Directorate:** Responsible for a decennial census of population and housing for the United States, Puerto Rico and the Island Areas to fulfill constitutional, legislative and programmatic mandates; the American Community Survey, a soon to be nationwide, annual demographic survey.
>
> **Demographic Directorate:** Responsible for numerous demographic surveys to profile and measure the people, housing and institutions of the Nation; compile and release annual population and housing estimates and population projections.
>
> **Economic Directorate:** Responsible for a quinquennial economic census for the United States, Puerto Rico and the Island Areas as well as numerous annual, quarterly and monthly surveys designed to measure the Nation's economy and provide information on federal, state and local governments.

1.3 Services

The Office provides tabulation and dissemination services. **Tabulation services** refer to the activities related to aggregating data collected on individual responses to a survey or census into summarized statistical data suitable for public release. **Dissemination services** refer to the activities related to internal staging and release of approved statistics to the Internet. These services also include the provision of interim results for other areas of the Census Bureau to process and release in different media and formats, such as, CD-ROM and DVD, File Transfer Protocol (FTP) or Adobe Acrobat (PDF) for printed publications.

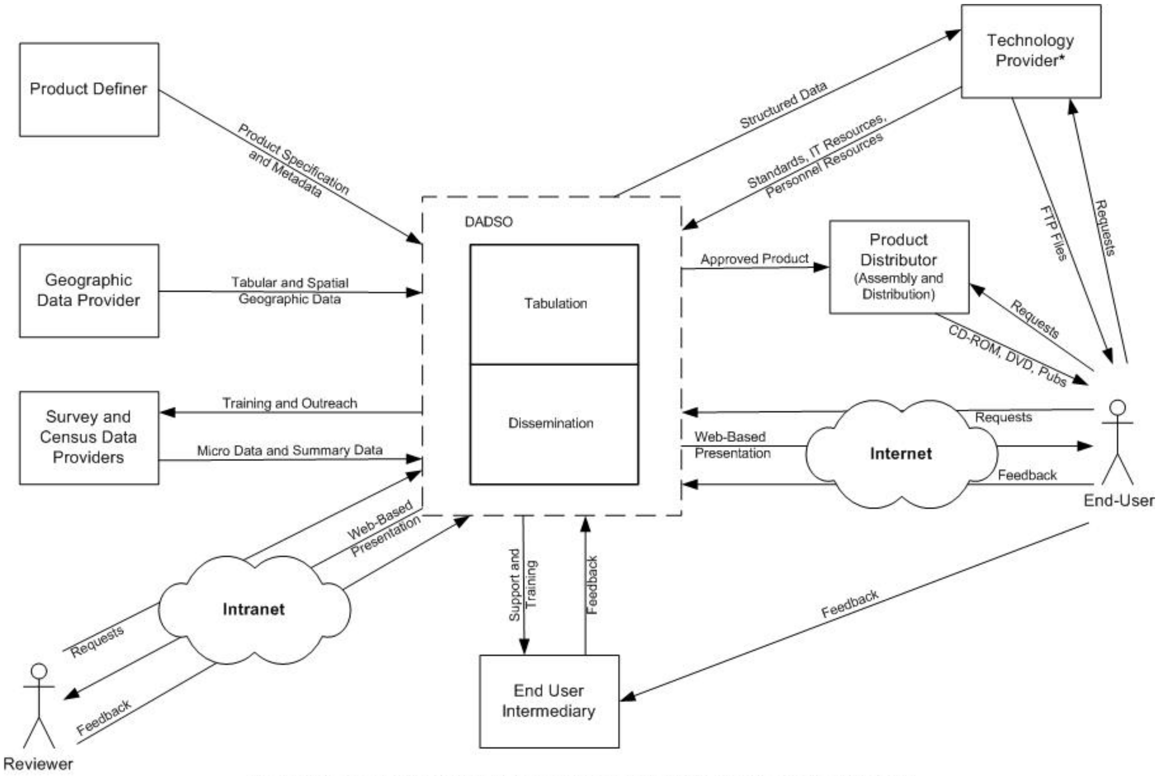
Figure 1 provides a high-level view of the Office, its services, and its inputs and outputs.

The **Survey and Census Data Provider** (lower left) can be of any of the three directorates described in Table 1 below. Each provides inputs of either microdata or summary data, and the associated metadata. These files along with inputs from the **Geographic Data Provider** and the **Product Definer** are used in the tabulation and dissemination of various data products.

The **Geographic Data Provider** provides both tabular and spatial geographic data. Tabular data are used to drive the tabulation process, as well as to complete dissemination products. Spatial data are used for map products.

.

[2] See http://www.census.gov/main/www/m-img/orgchart.jpg for the complete organizational hierarchy of the Census Bureau.



* Currently, hardware resources and operation of those resources for the Office's tabulation and dissemination activities are provided directly to the Data Access and Dissemination Office under the terms of the DADS contract.

Figure 1: Context Diagram

The **Product Definer** represents the subject matter experts who provide detailed specifications for the content, structure, presentation and dissemination method for each product that is tabulated or disseminated. Final product design is the result of an ongoing dialogue and consultation with stakeholders and end users of the data.

The **Reviewer** verifies that products are built according to specification, and reviews and clears them for public release. The review is conducted on internal Intranet systems in a secure environment by the same subject matter experts who define the products.

The **Product Distributor** assembles the approved products for distribution through the various media: Internet, CD-ROM and DVD, FTP and Adobe Acrobat (PDF). The Office is primarily responsible for the Internet distribution of products and data.

The **End User Intermediary** is the first line of contact for end users. These intermediaries may be call centers, regional offices, and support staff for a wide network of secondary distributors of Census Bureau data and information. The Office provides training and technical support to each. They, in turn, interact with end users of the data.

The **Technology Provider** defines information technology standards and provides personnel and infrastructure resources to support automated dissemination.

Table 1 provides a summary of the **Survey and Census Data Providers** supported by the Office. In the table, program refers to an established, ongoing data collection activity such as a survey, census or estimate program. Note that this is not a complete list of the data providers or programs of the Census Bureau.

Table 1: Survey and Census Data Providers

Directorate	Program	Years Supported	Tabulation Services	Dissemination Services
Decennial	Decennial Census	1990 – United States	No	Yes
		2000 – United States and Puerto Rico	Yes	Yes
		2000 – Island Areas	No	Yes
		2000 – Puerto Rico (Spanish)	No	Yes
	American Community Survey	1996 forward – United States	No	Yes
Demographic	Population Estimates	2001 forward – United States and Puerto Rico	No	Yes (part)
	School District Data File	2000 – United States	Yes	No
Economic	Economic Census	1997 – United States	No	Yes (part)
		2002 – United States	No	Yes
		2002 – Puerto Rico and Island Areas	No	Yes
	Non-Employer Statistics	2002 – United States	No	Yes
	Survey of Business Owners	2002 – United States	No	Yes
	Business Expenditure Survey	2002 – United States	No	Yes

Appendix A contains a timeline of the release of the various products for each Directorate and Program. More details on the types of products released by the Office are provided in Section 3.2.

1.4 Internal Partnerships

Successful implementation of tabulation and dissemination services depends upon strong partnerships with other key areas of the Census Bureau. These partnerships include extensive, ongoing communication, as well as the receipt of various inputs and provision of services. Table 2 lists the directorates that provide services to the DADS Office. The "Roles" supported by the directorates are based on the categories defined in section 1.3. The "Major Inputs" column lists the data or services provided by the directorate to the DADS office. The "Major Outputs" column lists the data or services provided by the

DADS office in support of that directorate. Outputs are delivered either to the supported directorate or to the end-user community on behalf of the supported directorate.

For example, training and outreach support is provided by the Office to various divisions and offices throughout the Census Bureau that interact with the public and other end users of the published data. The Office also provides intermediate outputs to other areas of the Census Bureau supporting other dissemination media.

Table 2: Partners in Tabulation and Dissemination

Directorate	Roles	Major Inputs	Major Outputs
Communications	• End User Intermediary	• Guidance and direction on dissemination activities targeted to Congress and the media	• Staff training and support for outreach activities for the media and general public • Web development services and content updates
Decennial Census	• Geographic Data Provider • Product Definer • Survey and Census Data Provider	• Geographic data to drive tabulation process, and to complete table and map products • Coordination and management of the Decennial products development and release • Microdata files and summary files	• Tabulation and dissemination services for the Decennial Census • Dissemination services for the American Community Survey
Demographic	• Product Definer • Reviewer • Survey Data Provider	• Product specifications and metadata • Product review and clearance • Subject matter expertise • Summary data	• Dissemination services for Population Estimates • Training of end user call center staff within Demographic Directorate and support of outreach activities • Specialized data analysis
Economic	• Product Definer • Survey and Census Data Provider	• Product specifications and metadata • Summary data	• Dissemination services for Economic Censuses and Surveys
Field Operations	• End User Intermediary	• Guidance and direction on dissemination activities and product design	• Training and outreach activities to the 12 regional offices
Finance and Administrations	• Product Distributor		• Outputs for creation of CD-ROM/DVDs and printed publications

Directorate	Roles	Major Inputs	Major Outputs
Marketing and Customer Liaison	• End User Intermediary	• Annual Internet end user satisfaction survey • Guidance and direction on dissemination activities and product design	• Training to call center staff within Marketing and Customer Liaison Directorate and support of outreach activities • Web development services and content updates
Methodology and Standards	• Technology Provider	• Standards for on-line presentation of statistical data and products • Resources for Section 508 testing • Resources for system usability testing	
Information Technology	• Technology Provider	• Guidance and standards for systems hardware, software and security • Internet dissemination standards • Secure hosting for all hardware • Telecommunications services	• File outputs for staging on an Bureau wide FTP server

2 Tabulation

The Office today provides tabulation services for two directorates, the Decennial Directorate and the Demographic Directorate. The Decennial Directorate relies on the Demographic Directorate for product design, subject matter expertise and technical direction (See Table 2). For the decennial cycle in 2000, the Office created all major data products for the Census 2000. It will be responsible for the same task for the decennial census in 2010. In addition, the Office created a special tabulation[3] based upon Census 2000 data, the School District Data File, for the Demographic Directorate.

2.1 Tabulation Inputs

In 2000, the decennial census employed a short-form questionnaire[4] to collect information about every housing unit and group quarters and every person in the United States, Puerto Rico and the Island Areas. The output of that effort was a microdata file, the Hundred Percent File. In addition, the census employed a long-form questionnaire[5] to collect information from a sample of housing units and group quarters. The output of that effort was a second microdata file, the Sample File. Table 3 below provides some basic

[3] Special tabulations are not standard Census Bureau data products and are generally sponsored by other public or private organizations. A list of special tabulations is available at http://www.census.gov/mp/www/spectab/specialtab.html.
[4] See http://www.census.gov/dmd/www/pdf/d61a.pdf.
[5] See http://www.census.gov/dmd/www/pdf/d02p.pdf.

and approximate information on the sizes of these microdata files. In 2010, the American Community Survey will replace the long-form survey.

Table 3: Census 2000 Microdata Sources

Microdata Source		Number of Records	Number of Fields
Hundred Percent File	Block	8,262,363	63
	Housing Unit	117,323,117	40
	Person in Housing Unit	277,405,109	77
	Group Quarters	192,286	7
	Person in Group Quarters	7,825,407	88
Sample File	Block	8,262,363	63
	Housing Unit	18,566,660	152
	Person in Housing Unit	43,166,083	234
	Group Quarters	138,604	15
	Person in Group Quarters	880,579	234

The Office received these two microdata files and geographic input files from the Decennial Directorate. From these inputs, the Office produced a collection of summary files containing aggregate data (counts, sums, means, medians, ratios and other aggregate measures) structured in a tabular format. In all, the Office produced eight major summary file products for Census 2000. Tables 4 and 5 contain basic but approximate information about each of these outputs. Extensive documentation for each summary file is available in American FactFinder.[6]

In most instances, tabulations were performed on a state-by-state basis with a subsequent effort to prepare results for geographic areas (such as regions, divisions and the United States itself) that cut across state boundaries. Depending on the size and complexity of the product, the computer executed tabulations for a single summary file may have taken two to five months. For example, Summary File 2 and Summary File 4 required a longer production and release schedule in order to produce tables not only for the total population, but also for each of 249 or 335 population groups, respectively.

2.2 Tabulation Outputs

The data provided in Table 4 suggest the enormous volume of the outputs of the Census 2000 Decennial tabulation process. The "Geographies" column lists the number of distinct spatial units over which data are tabulated (e.g. state, county, place). The "Iterations" column lists the number of population groups over which data is tabulated (e.g. Asian, Apache). The Census 2000 product contains a table associated with each base table, each geography, and each iteration. Hence, multiplying the values in the three columns provides an upper bound on the number of tables that are created for the

[6] See http://factfinder.census.gov/servlet/MetadataBrowserServlet?type=all&id=program&_lang=en for links to technical documentation. See http://factfinder.census.gov/servlet/DatasetMainPageServlet?_program=DEC&_lang=en&_ts= for definitions of the products.

product. The actual number of tables disseminated with the product may be less then this upper bound because of disclosure rules which exclude data derived from small sample populations. These disclosure rules are intended to prevent the release of confidential information. .

Table 4: Characteristics of Tabulation Outputs

Microdata Source	Census 2000 Product	Geographies	Base Tables	Table Cells	Iterations
Hundred Percent File	Redistricting Data Summary File	9,951,705	4	288	1
	Summary File 1	9,891,699	286	8,125	1
	Summary File 2	783,774	47	825	250
	108th Congressional District 100% Summary File	156,440	286	8,113	1
Sample File	Summary File 3	1,925,001	813	16,534	1
	Summary File 4	837,407	323	7,880	336
	108th Congressional District Sample Summary File	158,440	813	16,534	1
	American Indian and Alaskan Native Summary File	862	323	7,880	1,257

Table 5 provides an estimate of the number of months required to produce each of the Census 2000 products. This table also contains an estimate of the number of files which are included in each product. A file is defined as a collection of data, usually consisting of several cross-tabulations and tables.

Table 5: Release of Tabulation Outputs

Census 2000 Product	Number of Files	Production Start	Months in Production
Redistricting Data Summary File	156	Jan-01	1
Summary File 1	2,100	May-01	3
Summary File 2	33,142	Nov-01	4
Summary File 3	4,081	Jul-02	1
Summary File 4	358,314	Apr-03	2
108th Congressional District 100% Summary File	2,080	Dec-02	1
108th Congressional District Sample Summary File	4,004	Dec-02	2
American Indian and Alaskan Native Summary File	41,231	Jul-03	2

2.3 Tabulation Processing

The Office developed the Data Product Production system to perform the tabulations required by Census 2000. The system was built around a commercial, off-the-shelf software package, SuperSTAR from Space-Time Research, which served as the core tabulation engine. Although the system was used for all tabulation products, the timing of receipt of detailed product specifications and requirements unique to each product typically required system modifications prior to each major tabulation initiative. Quality requirements dictated strict configuration management controls and extensive testing. Outputs were examined both by automated processes and analysts who manually reviewed the data using an internal, protected version of American FactFinder.

2.3.1 Unique Challenges

The summary file products produced for Census 2000 are complex products. The technical documentation for Summary File 3 alone runs to more than 1250 pages.[7] Among the greatest challenges of the tabulation undertaking are the complexity of geographic hierarchies, the uniqueness of some statistical measures that must be applied, the volume and complexity of the data that must be managed in processing, the overlay of techniques to protect the confidentiality of census respondents, and the demands of quality assurance.

Within each summary file product, a predefined set of counts, sums, means, medians and other aggregate values must be computed for some subset of the more than 10 million geographic areas for which the decennial census prepares results. These geographic areas are related hierarchically and there are numerous independent hierarchies[8]. Although most of the aggregate calculations made in the course of preparing a summary file involve commonly used statistical techniques, some involve techniques rarely used outside the Census Bureau. The challenges posed by these exceptions can be considerable.

The sheer number of calculations that must be performed and the management of those unique results contribute substantially to the difficulty of the tabulation effort. Frequently, the complexity of the result may require the preparation of many intermediate products. For example, more than 17 million intermediate files were produced to complete the tabulation of Summary File 4. The sheer volume of data that must be managed proves a continual challenge.

The Census Bureau demands 100% accuracy in its tabulation activities. Efforts to achieve this objective require extensive quality assurance activity. Significant components of both system and intermediate products were created specifically and only for the purposes of quality control.

[7] The technical documentation for Summary File 3 can be found at http://www.census.gov/prod/cen2000/doc/sf3.pdf.

[8] A description of geographies can be found at http://www.census.gov/geo/www/reference.html

2.3.2 Title 13 Protection

Title 13 of the United States Code protects all data collected by the Census Bureau and guarantees the confidentiality of census information. There are strict rules for the handling of data and penalties for any disclosure. In addition, the Census Bureau uses statistical methods throughout the processing cycle to guarantee confidentiality of the data. Among the techniques applied to Census 2000 products were substitution of a specified fixed value for any result that might reveal confidential information, rounding of sums and rule-based suppression of results for individual geographies.

3 Dissemination

The Office disseminates data and information for the programs it supports (see Table 1) over the Internet[9] from its site at http://factfinder.census.gov/. For some of these same programs, the Office also provides intermediate products to other organizations within the Census Bureau for assembly and publication via other media, both electronic and print.

3.1 End Users

Achievement of Census Bureau strategic goals requires a continuous improvement process to meet the ever-changing needs of its end users by enhancing data products, services and dissemination. In response, the Office routinely monitors satisfaction with its services through an online Internet survey, attendance at stakeholder meetings, system feedback and email, and contacts with various end user call centers throughout the country. The Office also analyzes Internet survey results to categorize and better understand end users.

3.1.1 End User Segmentation

End users of Census Bureau data and information range from school age children to social scientists. Their questions range from the simplest, such as “What is the population of my town?” to requests for detailed time trend analyses to support policy decisions. In the past, end users were typically categorized according to occupation, such as government, businesses, libraries, and academia. With the advent of Internet dissemination, the Office now segments users based on behavioral and other criteria including:

- types of queries or data products requested, and
- prior knowledge and understanding of Census Bureau data and products.

The end user segmentation shown in Table 6 is based on end user interviews and results of previous surveys of Internet usage. It delineates four basic end user segments, Surfers, Portrayers, Manipulators, and Extractors.

[9] Other areas of the Census Bureau manage the balance of Internet dissemination activities. The central site for dissemination by the Census Bureau is http://www.census.gov/

Table 6: End User Segmentation

Segment	Priority Goal	Knowledge of Products and Data	Relevance of Data to Job	Computer Expertise
Surfers	Quickly understand purpose and function and content of the site	Little to none	None	Varies
Portrayers	Quick answers to basic questions	Minimal to knowledgeable	Need data for work/study	Medium/Low
Manipulators	Perform queries and cross tabulations	Experienced	Major component	High
Extractors	Download/extract data for further analysis	Experienced to expert	Primary component	High

The Office has adopted an approach of providing services and products targeted towards each segment. This means organizing content and developing functions that satisfy the various needs and goals of each segment.

3.1.2 End User Support

The Office supports end users in a number of ways:

Feedback: Each public system allows end users to send in questions and comments. The Office answers these questions and uses them to monitor problems and issues with the system. Suggestions for new functions and corrections to the system are made based on this feedback.

Training: The Office provides training for internal liaison staff to learn to use each public system. These staff in turn train and educate the public on how to use the dissemination systems and products. Ongoing support in the form of briefings and seminars are provided to internal staff as the systems are updated and new major products are released. Release notes and product announcements are also placed online for the general public access.

Tutorials: Online tutorials and user’s guides have been developed to facilitate distance learning.

The public expects the public online dissemination systems provided by the Office to be available 24 hours a day, 7 days a week unless otherwise announced. Every effort is made to meet that expectation. System maintenance is scheduled to avoid core business hours whenever possible.

3.2 Dissemination Products

The Census Bureau disseminates a spectrum of products providing data and information to end users. Among the more important of these products are data files, tables, maps, analytical reports and documentation.

3.2.1 Data Files

Data files are collections of data designed for computer manipulation. They are typically accompanied by extensive documentation or metadata detailing the content and format of the files and they are frequently used to make large amounts of data available to end users.

The Census 2000 Summary File 1 is actually a set of 2,100 data files. The content of Summary File 1 includes 286 cross-tabulations and tables of derived measures prepared for each of approximately 10 million geographies, divided into state and national components and further subdivided into files of manageable size.

3.2.2 Tables

Tables are collections of data systematically arranged in rows and columns with appropriate titles, column headers, row labels and other text to fully identify the table and each value. Tables are generally designed for viewing. Subject matter experts define the content, structure and format of each table.

Numerous tabular products, including profiles and geographic comparison tables, are not the direct result of the tabulation but are constructed using data extracted from other tabular products. Although users may be unaware of the distinction, the **base tables** that result from tabulation are used to construct **derived tables** that may further aggregate or modify the presentation of that data. The tables of the Census 2000 Summary File 1 are considered base tables, while the profiles, quick tables, and geographic comparison tables associated with the same summary file are all derived from the 286 base tables of Summary File 1.

3.2.3 Maps

Like tables, maps are products meant for viewing. **Reference maps** graphically describe the bounds of geographies used in tabulation. They typically include orienting features such as roads and water bodies in addition to geographic boundaries. **Thematic maps** help users visualize geographic patterns in data. The most commonly used thematic map is the choropleth map, in which each geographic area is colored or shaded based on an associated data value. Examples of reference and thematic maps can be found in American FactFinder.

3.2.4 Analytical Reports

Analytical reports are typically published reports containing narrative description with supporting tables, maps and charts providing information on a particular topic. They may appear in both print and electronic media. The electronic versions of these reports may be disseminated over the Internet. The Reference Shelf in American FactFinder provides links to numerous analytical reports. The Census 2000 Briefs series provides many excellent examples.

3.2.5 Documentation and Metadata

Documentation and metadata provide information to clarify the meaning of data. Each may provide information on data collection instruments, data processing procedures, data accuracy, the meaning of terms and concepts, the interpretation of coding schemes, and a host of similar topics. The technical documentation for major products such as Census 2000 Summary File 1 are typically published documents made available in electronic form. Metadata is structured and stored as data. An example of metadata can be seen by clicking the hyperlinked title of any table in American FactFinder.

3.3 Dissemination Systems

This section describes the systems that the Office currently maintains to provide dissemination services.

3.3.1 American FactFinder System

The American FactFinder system was developed and is maintained to provide Internet dissemination of Census Bureau data and information. The system supports three distinct web sites, American FactFinder, FastFacts for Congress, and American Indian and Alaskan Native Data and Links. The system allows Internet users to locate, select and view or download data files, tables, maps, analytical reports and documentation from each of these three web sites. The system delivers results in English or Spanish.

Figure 2 details the application architecture of the American FactFinder system. The system employs a three-tier architecture. Web servers running IBM HTTP Server and supported by an IBM Edge Server (Network Dispatcher) load balancer comprise the first tier. Application middleware, hosted in an IBM WebSphere Application Server container, consisting of custom Java and ESRI ArcIMS (Internet Mapping Server) software, comprises the second tier. The third tier provides data, metadata and spatial data using Oracle 9i and ESRI ArcSDE (Spatial Data Engine).

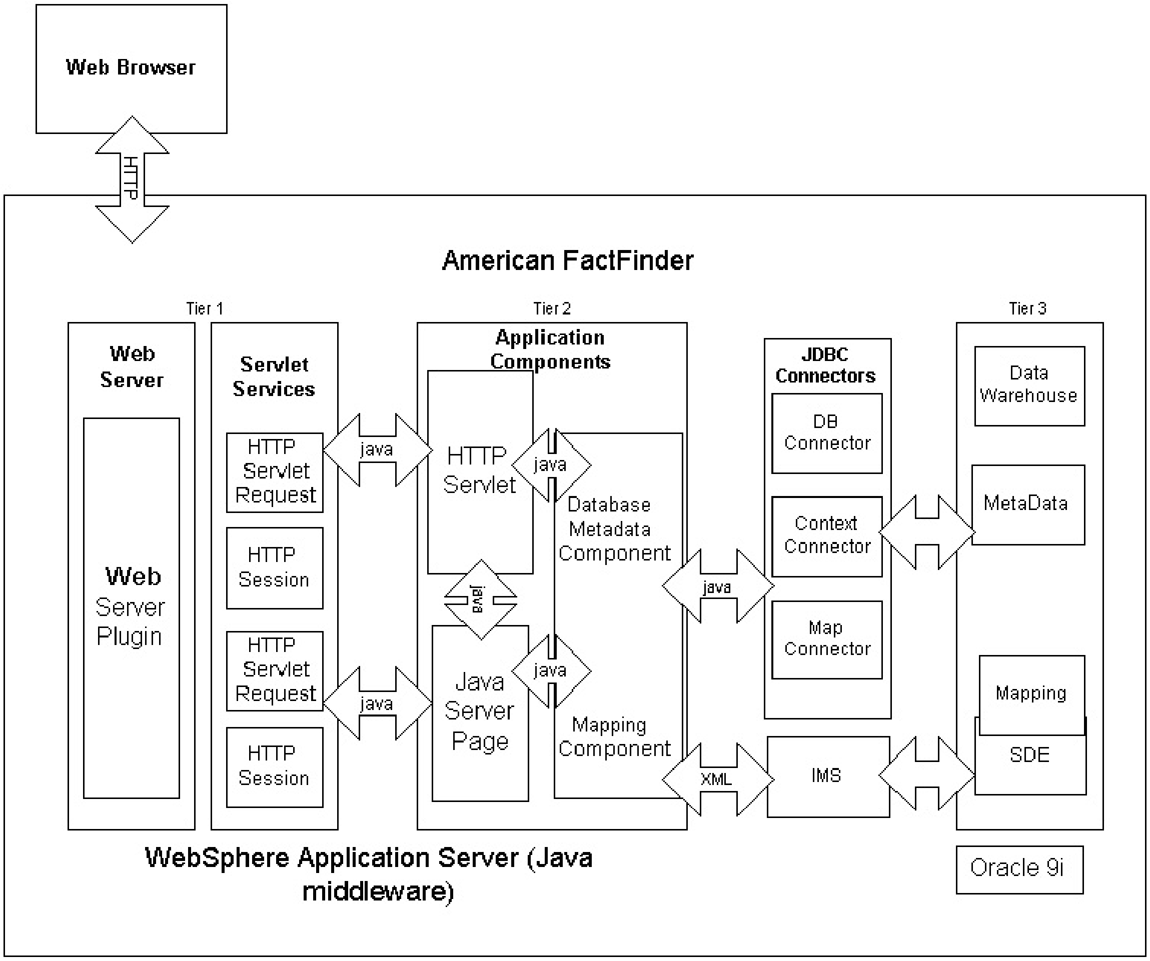


Figure 2: American FactFinder System

The American FactFinder[10] web site draws upon all of the data and information within the American FactFinder data warehouse and provides links to related resources available from other Census Bureau providers as well as other members of the federal statistical community. Table 7 provides information on the number of products maintained in the American FactFinder data warehouse. The warehouse also includes more than 500 layers of spatial data covering the United States, Puerto Rico and the Island Areas.

Table 7: Number of Tables and Maps in American FactFinder

Directorate	Program	Years Supported	Number[11]			
			Data Sets	Base Tables	Derived Tables	Map Themes
Decennial	Decennial Census	1990 – United States	2	402	8	134
		2000 – United States and Puerto Rico	8	2903	1178	618

[10] See http://factfinder.census.gov/

[11] Represents the number of unique base tables, derived tables and map themes released through American FactFinder. Each table or map can be produced for a large number of geographic areas.

		2000 – Island Areas	4	1684	0	0
		2000 – Puerto Rico (Spanish)	4	1760	978	287
	American Community Survey	1996 forward – United States	16	6966	4	0
Demographic	Population Estimates	2001 forward – United States and Puerto Rico	3	6	31	12
Economic	Economic Census	1997 – United States	332	332	5	85
		2002 – United States	719	719	11	TBD
		2002 – Puerto Rico and Island Areas				
	Non-Employer Statistics	2002 – United States				
	Survey of Business Owners	2002 – United States				
	Business Expenditure Survey	2002 – United States				

FastFacts for Congress[12] was launched at the start of the 108th Congress. It focuses on congressional districts and is meant to meet the special needs of Congress. The American Indian and Alaskan Native Data and Links[13] web site focuses on the user community interested in data and information on American Indians and Alaskan Natives. This web site consists of only a handful of static pages providing links to other resources. Unlike FastFacts for Congress, links to American FactFinder tables and maps on the American Indian and Alaskan Native Data and Links web site merely take the user to the American FactFinder web site rather than providing a parallel, customized access to data resources.

American FactFinder developers use a custom Virtual Table Generator, a desktop tool, to create and review derived table shells. The Generator accepts format and sourcing specifications for a derived table and creates a table shell showing the table columns, rows, text, and rules for populating each derived table cell from base table data. Once this intermediate output has been reviewed, corrected, and approved, the Generator outputs metadata files used by the American FactFinder system to create derived tables.

A Printed Report PDF Generator allows developers to use the American FactFinder system to create tabular output that meets publication quality design standards in PDF format. This tool uses an application maintained by Fenestra Technologies Corporation. Custom code takes derived tabular data from American FactFinder together with layout formatting information and produces an XML output file. The Fenestra application takes the XML file together with additional paper-specific formatting information and creates a publication quality PDF file. The Generator was used to produce thousands of pages of tables for Census 2000 printed reports. The PDF outputs were delivered to an integrator for further document assembly before being delivered to the printer.

3.3.1.1 Environments

Operation of the American FactFinder System requires multiple environments allowing different work efforts to move forward in parallel. Each environment has a unique

[12] See http://fastfacts.census.gov/home/cws/main.html

[13] See http://factfinder.census.gov/home/aian/index.html.

combination of data store, metadata store, spatial data store, and software release. Many of the environments exist only “virtually”, with a given component (e.g., development metadata store) being used by more than one environment. Following is a brief description of each American FactFinder environment:

> **Development:** Dedicated to the coding and unit testing phases of American FactFinder software development.
>
> **Product Assurance:** Dedicated to the system testing phase of American FactFinder software development.
>
> **Internal Review:** For product definer review of data and its presentation prior to data deployment and/or software release.
>
> **External Review:** An area of the external system, hidden from public view, used to verify success of deployment steps as they are executed in real-time.
>
> **External Production:** The publicly available environment for all American FactFinder system web sites.

3.3.1.2 Release Schedule

For the American FactFinder system, a typical year includes two to three major software releases. Release timing and content is largely driven by the functional needs of new data deployments, along with current commitments to improve general functionality. For some data deployments, no software release is required, because the existing functionality satisfies all requirements. Additionally, limited updates to static content do not require a software release.

3.3.1.3 Data Deployments

The schedule of American FactFinder data deployments is highly variable, reflecting the fluctuating rate at which data tabulations are produced by the various programs. In 2003, eight data sets were released on American FactFinder with several requiring phased releases resulting in 23 separate deployments.[14] Phases of release for a single data set may span months. Some typical release strategies include:

> **By geographic area:** For Census 2000 Summary File 4, data for several states were released each week, with nation-level geographies crossing state lines included in the final week's release.
>
> **By geographic type:** For the annual Population Estimates Program data set, data were first released for the nation and for states, next for counties, and finally for cities and towns.

[14] See Appendix A for a timeline of data deployments.

By table: For the annual American Community Survey data set, data are first released for core tables covering basic subject characteristics, and several months later for non-core tables that cover the same subject characteristics but iterated for major race groups.

3.3.2 Advanced Query System[15]

The Advanced Query system allows Internet users to construct queries based on data available in Census 2000 microdata files. The resulting queries must be compliant with strict rules for query construction that serve as part of a series of restrictions guarding the confidentiality of the underlying data. Compliant queries are run against Census 2000 microdata files, returning results to the application. If the results can pass a series of stringent filters, they are returned to the users who requested them. Query processing and results filtering are accomplished in real time. Information on the methodology used to protect against disclosure in Advanced Query is detailed in the soon to be published paper *American FactFinder: Disclosure Limitation for the Advanced Query System*.

Advanced Query is a MicroStrategy application working with a highly optimized deployment of the Census 2000 microdata in an IBM DB2 database. Windows 2000 based Microstrategy Web Servers on the Internet communicate across a security firewall with Microstrategy Intelligence Servers deployed in a protected zone. The Intelligence Servers in turn communicate across a second firewall with database servers deployed in a separate protected zone. Access to Advanced Query is password protected.

In 2004, Advanced Query remains under evaluation. The system is currently available only to users assisting with that evaluation. End users are supported by training, extensive documentation including guides and tutorials, and response to feedback. Updates and improvements to the system occur on an occasional basis.

4 Infrastructure

4.1 Data Product Production System Infrastructure

The Data Product Production system comprises two distinct hardware subsystems dedicated to development and production respectively. The infrastructure is detailed in Figure 3.

- The Data Product Production **development** system consists of an IBM S80 server with 8 CPUs and 16 gigabytes RAM.

- The Data Product Production **production** system consists of an IBM M80 server with 24 CPUs and 96 gigabytes RAM. At the height of decennial processing, two additional servers were needed to process the data. The additional machines were

[15] See http://advancedquery.census.gov/Login.asp.

IBM p680 unix servers with 24 processors and 96 GB RAM which were attached to a total of 18 TB of ESS storage.

The two Data Product Production systems share a 5.7 terabyte IBM Enterprise Storage Server system. All systems run under the IBM AIX operating system.

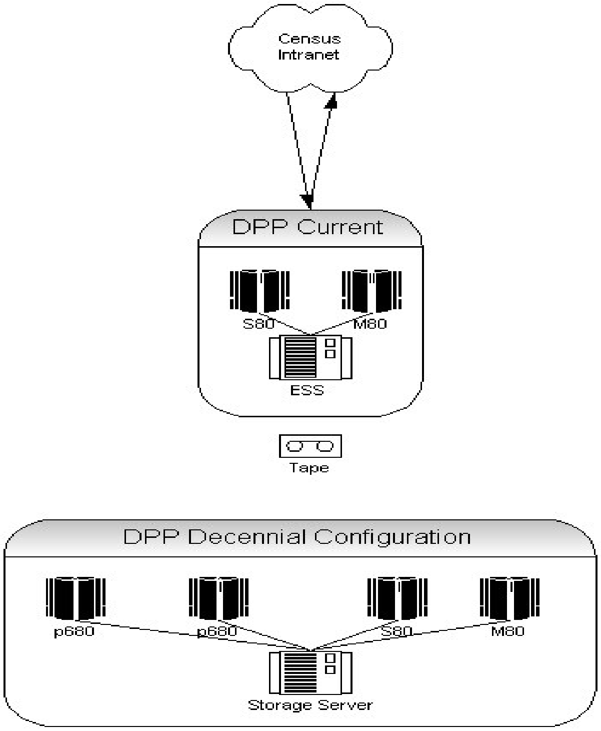


Figure 3: Data Product Production System Hardware Diagram

4.2 American FactFinder System Infrastructure

The American FactFinder system comprises three distinct hardware subsystems dedicated to development, internal review and production respectively. The infrastructure is detailed in Figure 4.

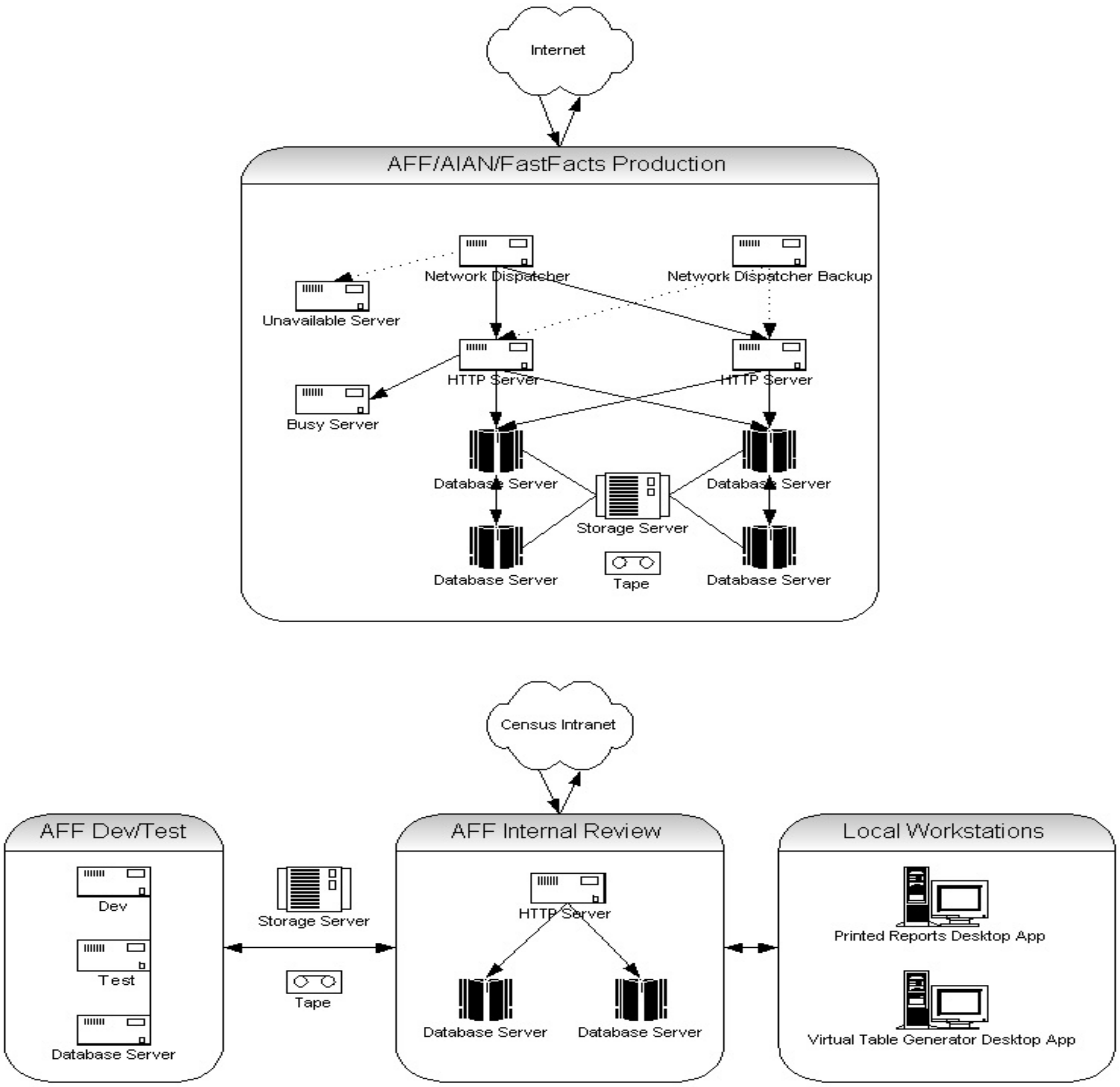


Figure 4: American FactFinder System Hardware Diagram

- The American Factfinder **development** system is accessible on the Census Bureau intranet and is used to develop and test new application code against test data and metadata. This subsystem consists of one IBM SP system with a control workstation and a single node and a second IBM SP system with a control workstation and three p630 nodes. Each node has 2 CPUs and 4 gigabytes RAM.

- The American Factfinder **internal review** system is accessible on the Census Bureau intranet and is used both for testing and reviewing Census Bureau data and metadata before it is released to the public. This subsystem consists of a three node IBM SP with a control workstation, one p630 node, and two p680 nodes. The p630 node contains 2 CPUs and 4 gigabytes RAM. The two p680 nodes have 6 CPUs and 20 gigabytes RAM each.

- The American Factfinder **production** system is accessible on the Internet and is used for public dissemination of products and information. This subsystem consists of four stand-alone systems and one IBM SP system. Each stand-alone system consists of one p615 node; these four systems sit in front of the IBM SP system. The IBM SP system consists of one control workstation and four p680 nodes. Two of the p680 nodes have 12 CPUs and 40 gigabytes of RAM; the other two p680 nodes have 24 CPUs and 96 gigabytes of RAM. Behind the system are two IBM RS/6000 servers for additional web services when the American FactFinder system is either busy or unavailable.

All of the SP systems run the IBM AIX operating system. There are two IBM Enterprise Storage Servers (ESS). The production ESS contains 13 terabytes of storage. The development and internal review systems share a second ESS containing 5.7 terabytes of storage.

The virtual table generator runs on a Novell shared network drive.

The printed report PDF generator runs on a Dell XEON 2.00 GHz processor with 3 gigabytes RAM. A backup machine is configured with the same characteristics and is used if the primary machine is unavailable.

4.3 Advanced Query System Infrastructure

The Advanced Query system comprises three distinct hardware subsystems dedicated to development, internal production and production respectively. The infrastructure is detailed in Figure 5.

- The Advanced Query **development** system consists of two IBM NetFinity servers. One server is used for web services and the other as an intelligence server. One IBM 6H1 server with 4 CPUs and 16 gigabytes of RAM is used for the backend IBM DB2 RDBMS database server.

- The Advanced Query **internal production** system is accessible on the Census intranet and uses one NetFinity server as a web server and one NetFinity server as an intelligence server. Both the IBM 6H1 and the IBM p680 servers, listed above, serve as the backend database server.

- The Advanced Query **production** system is accessible on the Internet and consists of three IBM NetFinity servers used as web servers. Three IBM NetFinity servers are used as intelligence servers. One IBM p680 server with 24 CPUs and 96 gigabytes RAM running IBM DB2 RDBMS acts as the backend database server. All storage is IBM Serial Storage Architecture (SSA) RAID 5 drives.

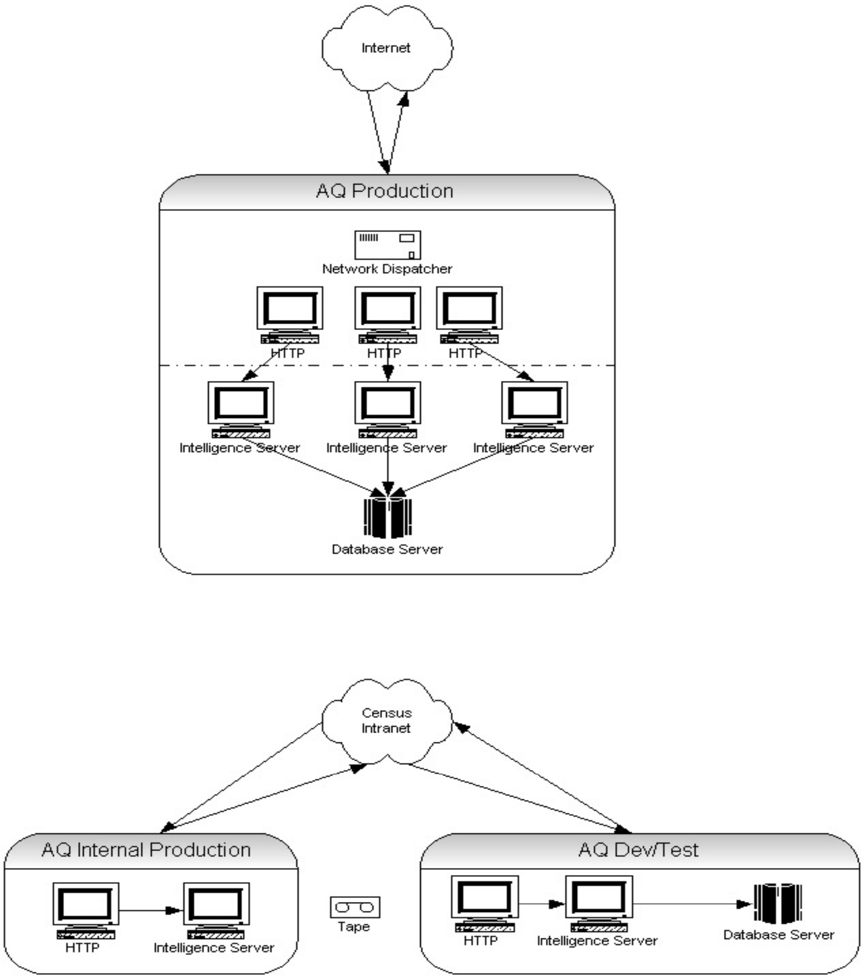


Figure 5: Advanced Query System Hardware Diagram

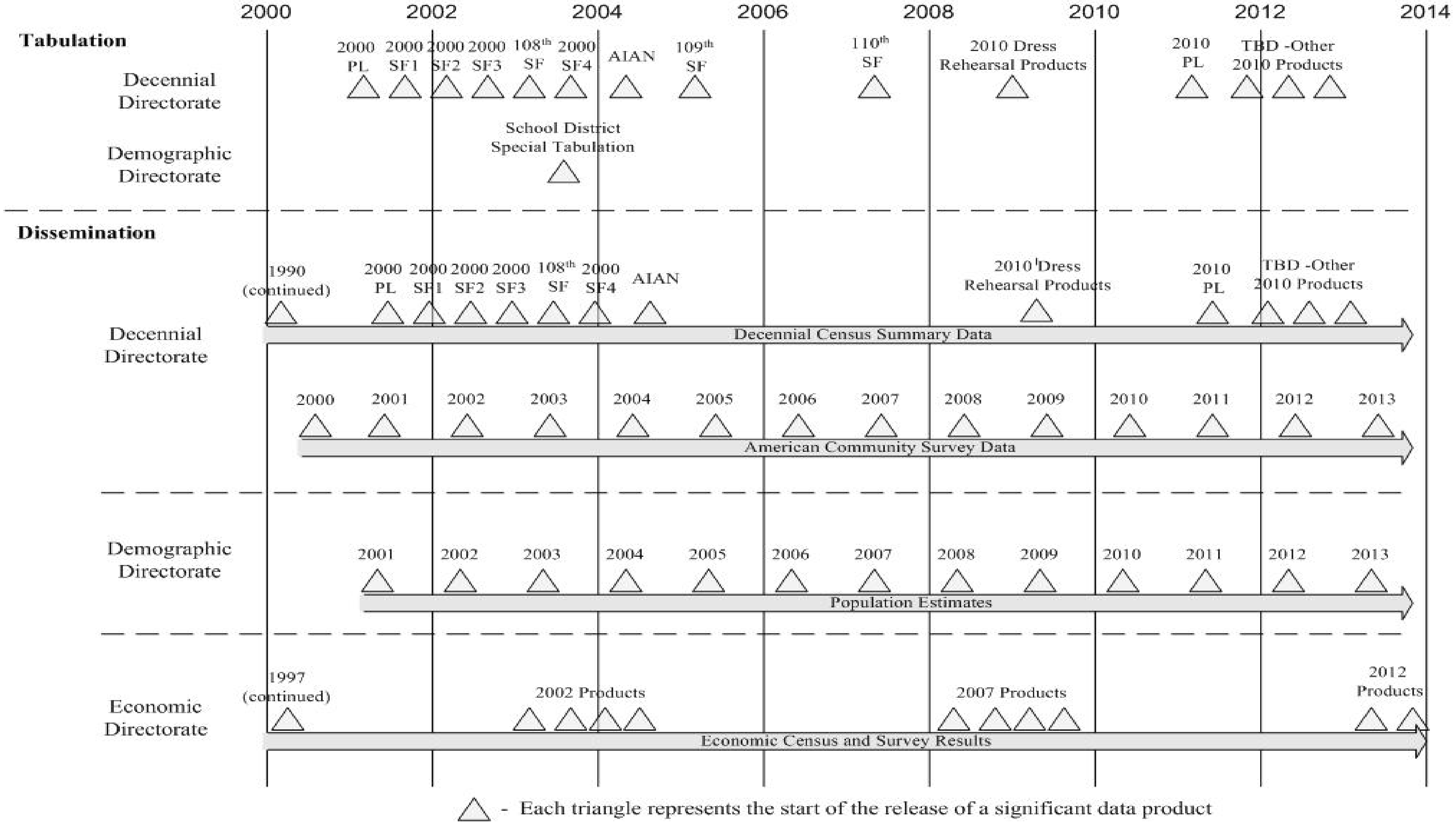
4.4 Networks

All systems are housed in the Bowie Computer Center in Bowie, Maryland and all are connected to the Census Bureau intranet. Systems connected to the Internet use a Census Bureau dedicated Internet service provider connection. Network devices and firewall hardware are provided by the Census Bureau Information Technology Directorate.

The SP systems use an IBM SP switch (150 MB – low latency – point to point) network running PSSP software. Virtual shared disk software, also part of the PSSP software suite, runs under the Oracle RDBMS and provides fault tolerant redundancy.

Development and operations are conducted from a single location in the Suitland Federal Center in Suitland, Maryland. High-speed networks connecting the Suitland Federal Center and the Bowie Computer Center serve as the link between the operations staff and the physical systems. For the most part, the hardware functions in a lights-out environment. Most of the interaction between operators and hardware occur when tapes are taken offsite for disaster recovery storage or when hardware upgrades or maintenance are required. All systems are backed up using IBM LTO and 3590 tape storage libraries. Separate storage libraries are maintained for systems on the Census Bureau intranet and systems on the Internet.

Appendix A: Timeline of Product Tabulation and Dissemination by Directorate



Appendix B: Glossary

Term	Definition
Census	A complete enumeration, usually of a population, but also of businesses and commercial establishments, farms, governments, and so forth.
Customer	Any Census Bureau directorate requiring services from the Data Access and Dissemination Office.
Data Provider	Any Census Bureau organization making Census Bureau data available to the Data Access and Dissemination Office for use in tabulation or dissemination activities.
Data Set	An abstraction used to refer to a collection of products. These collections are typically related to a single tabulation output.
Decennial Census	The census of population and housing, taken by the Census Bureau in years ending in 0 (zero). Article I of the Constitution requires that a census be taken every ten years for the purpose of reapportioning the United States House of Representatives.
Dissemination	The term is used broadly to refer to any activity following the completion of tabulation activities that contributes to putting Census Bureau data and information in the hands of end-users.
Economic Census	Collective name for the censuses of construction, manufactures, minerals, minority- and women-owned businesses, retail trade, service industries, transportation, and wholesale trade, conducted by the Census Bureau every five years, in years ending in 2 and 7.
End User	Anyone, either within or without the Census Bureau, who uses associated Census Bureau data and information.
End User Intermediary	The first line of contact for end users who need assistance in using and understanding dissemination products.
Geographic Data Provider	Internal partners who provide geographic tabular and spatial data and metadata to support tabulation and dissemination.
Geography	Any spatial unit used by the Census Bureau for data collection and tabulation. The spatial unit may be formed by legal or statistical boundaries. Examples are states, counties, places, county subdivisions, census tracts and census blocks. An instance of a spatial unit. For example, Alaska or Middlesex County, Connecticut.
Internet User	Any end-user who gains access to Census Bureau data and information using the Internet.
Intranet	The local and wide-area network maintained by the Census Bureau for internal use. In this paper, the term does not connote exclusive use of the HTTP protocol.
Island Areas	Refers to the following United States territories and protectorates: Virgin Islands of the United States, Guam, Commonwealth of the Northern Mariana Islands, and American Samoa.
Microdata	A collection of data reflecting responses taken from individual data collection instruments.
Population Estimate	Population estimates are basic population counts released between decennial censuses. Population estimates are released as of July 1 for each year.
Population Projection	Population projections are estimates of future counts of the resident population, families and households
Product	A term used broadly to refer to any Census Bureau sponsored output of Census Bureau data or information.
Product Definer	Any member of a team of individuals responsible for determining the

Term	Definition
	characteristics of a product, including content and format.
Product Distributor	A Census Bureau organization responsible for activities related to both the production and distribution of Census Bureau data and information for products requiring physical distribution, such as products made available on optical media or in print.
Product Specification	A document providing detailed description of the outputs of a tabulation or dissemination activity.
Program Sponsor	Any Census Bureau organization directly responsible for a census, survey, estimate or projection.
Reference Map	Maps that graphically describe the bounds of geographies used in tabulation.
Reviewer	Typically, a statistician or analyst responsible for examining the outputs of tabulation or dissemination activities for quality control purposes.
Summary Data	Summary data is a collection of data containing only aggregate representations (counts, sums, medians, means, etc.) of the contents of a microdata source.
Survey	A data collection activity involving observation or questionnaires for a sample of a population.
Survey and Census Data Provider	Internal customers who provide census and survey results for tabulation or dissemination.
Tabulation	A collection of activities required to produce summary data from microdata.
Thematic Map	Maps that are used to visualize geographic patterns in data by coloring or shading each geographic area based on an associated data value.
United States	The 50 states and the District of Columbia.